SPENDING A RAINY AFTERNOON WITH SAS - No. 2: ILLUSTRATIONS OF THE USE OF SAS IN FACILITATING CERTAIN STATISTICAL CALCULATIONS

by

William D. Bell

BU-542-M*

December 1974

Abstract

This mimeograph illustrates ways in which SAS (the Statistical Analysis System) may be used:

1. to facilitate computations required to make tests which depend on ranks;
2. to provide AØV's which include finer partitions of treatment and of residual sums of squares, on the original and/or on transformed data;
3. to make displays of residuals which may, at times, be useful in checking assumptions about the "error";
4. to provide Cox's (58) AØV which allows "for individual curvatures in one direction" in Latin squares;
5. to aid in making the calculation of the sum of squares associated with Tukey's (49) single degree of freedom for non-additivity;
6. to compute values of certain series of orthogonal polynomial functions of the (an) independent variable in a regression at values assumed by the independent variable when these values are unequally spaced and/or when it is desired to calculate the regression weighted according to a (known) function of the independent variable; the principle is illustrated by using it to provide Robson and Atkinson's (58) AØV.

---

* In the Biometrics Unit Mimeo Series, Cornell University, Ithaca, N. Y. 14853.

SPENDING A RAINY AFTERNOON WITH SAS - No. 2: ILLUSTRATIONS OF THE USE OF SAS IN FACILITATING CERTAIN STATISTICAL CALCULATIONS

by

William D. Bell

BU-542-M*

The data: We make use of two sets of data for purposes of illustration. One set was re-published in Bartlett (47). (The data was used in a course given in the Biometrics Unit by Professor W. T. Federer; it also appears elsewhere in the literature. Indeed, this must be one of Bartlett's familiar quotations.) Bartlett provides the data in the following tabular form:

Block						
A	(1)	(4)	(2)	(5)	(3)	(6)
	438	17	538	18	77	115
B	(3)	(2)	(6)	(1)	(5)	(4)
	61	422	57	442	26	31
C	(5)	(3)	(4)	(6)	(2)	(1)
	77	157	87	100	377	319
D	(2)	(1)	(5)	(3)	(4)	(6)
	315	380	20	52	16	45

Table 1

The numbers in parentheses are the treatment numbers, the type of herbicide applied to the plots, (1) being the control, no chemical applied. The numbers below the parenthesized ones give the number of poppy plants in a cereal plot.

---

* In the Biometrics Unit Mimeo Series, Cornell University, Ithaca, N. Y. 14853.

The other set of data was published in Cox (58). To quote Cox: "The following table shows mean weekly milk yields of four Guernsey cows starting at 17, 11, 15 and 7 weeks respectively, of their lactations in the first experimental week shown. Regarding these as uniformity trial data, a randomized 4x4 Latin square design has been imposed on the results ... ."

	Yields				Design			
	Cow				Cow			
Week	1	2	3	4	1	2	3	4
1	83.75	33.25	45.75	144.50	D	C	A	B
2	47.00	44.00	36.75	135.75	A	D	B	C
3	36.50	41.50	27.50	125.50	C	B	D	A
4	24.00	38.75	30.50	124.25	B	A	C	D

Table 2

Illustrations of calculations involving ranks: This section is mainly aimed at indicating how SAS can be employed to facilitate computations involved in making the Friedman test and also (effective simultaneously) the analogous median test, not to mention how calculations for other statistics associated with tests of this same class could also be effected more easily. If one assumes that there is no block-treatment interaction* then the distribution-free test of no treatment differences described in Hollander and Wolfe (73), and attributed by them to Friedman et al., may be appropriate. The statistic used to make the test is given by

---

* Although, as Dr. Federer (74) has pointed out, this assumption is apparently untenable.

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2\right] - 3n(k+1) \qquad (1)$$

where

$k \equiv$ no. of treatments ;

$n \equiv$ no. of blocks ;

and

$R_j = \sum_{i=1}^{n} r_{ij}$, r_{ij} being the rank

of the j^{th} observation within the i^{th} block .

Hollander and Wolfe provide instructions and a table for applying this test. Although the table does not appear to be as extensive as it ought to be, approximating the null distribution of S with the χ^2_{k-1} distribution would likely serve in this instance.

The corresponding median test statistic is given by Hájek and Sidák (67):

$$Q = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} A_{ji}\right)^2 - \tfrac{1}{4}n^2k \qquad (2)^*$$

where

$$A_{ji} = \left\{\begin{matrix} 1 \\ \frac{1}{2} \\ 0 \end{matrix}\right\} \quad \text{if } X_{ji} \text{ is } \left\{\begin{matrix} > \\ = \\ < \end{matrix}\right\} \text{ median within } j^{th} \text{ block} \quad .$$

---

* This statistic is also of χ^2 type.

```
01              DATA ONE; INPUT BLOCK $ 1 Y_1 2-4
02                                        Y_2 5-7
03                                        Y_3 8-10
04                                        Y_4 11-13
05                                        Y_5 14-16
06                                        Y_6 17-19;
07              TREAT=1; Y=Y_1; OUTPUT;
08              TREAT=2; Y=Y_2; OUTPUT;
09              TREAT=3; Y=Y_3; OUTPUT;
10              TREAT=4; Y=Y_4; OUTPUT;
11              TREAT=5; Y=Y_5; OUTPUT;
12              TREAT=6; Y=Y_6; OUTPUT;
13
14                MACRO MAC1 SET ONE; IF BLOCK= %
15                MACRO MAC2
16                           Y2=Y;
17                           Y3=-Y;
18                                 PROC RANK ; VAR Y;
19                                 PROC RANK  GROUPS=2; VAR Y2;
20                                 PROC RANK  GROUPS=2; VAR Y3; %
21                MACRO MAC3 Y4=AVG(Y2,MOD(Y3+1,2)); %
22
23              CARDS;
24              A438538 77 17 18115
25              B442422 61 31 26 57
26              C31937715/ 87 77100
27              D380315 52 16 20 45
28              DATA BLOCK_A; MAC1 'A';MAC2 DATA BLOCK_A2; SET BLOCK_A; MAC3
29              DATA BLOCK_B; MAC1 'B';MAC2 DATA BLOCK_B2; SET BLOCK_B; MAC3
30              DATA BLOCK_C; MAC1 'C';MAC2 DATA BLOCK_C2; SET BLOCK_C; MAC3
31              DATA BLOCK_D; MAC1 'D';MAC2 DATA BLOCK_D2; SET BLOCK_D; MAC3
32
33              DATA TWO; SET BLOCK_A2; SET BLOCK_B2; SET BLOCK_C2; SET BLOCK_D2;
34              PROC SORT; BY TREAT;
35              PROC MEANS; VAR Y Y4; BY TREAT;
```

Program explication: The four data cards represented by lines 24 through 27 of the listing are read by lines 1 through 13, and 22 and 23; SAS data set "one" is produced. The central principle of operation of these statements is explained in the SAS manual in the section describing the use of the output statement.

A very rudimentary macro processor is a part of SAS. Macro definitions in SAS begin with the characters "MACRØ" and end with the character "%". Thus, in this program, lines 14 through 21 constitute the definition of three macros called "MAC1", "MAC2", and "MAC3". Now any time after Line 21 that SAS encounters the character string "MAC1", SAS will replace it with the character string in the macro definition, starting with the characters "SET" and ending with "BLØCK=". Of course SAS will behave in a similar fashion with respect to the macros MAC2 and MAC3. Thus, the first line invoking these macros will be expanded by SAS to form:

```
data block_a;   set one; if block = 'a';
        y2 = y; y3 = -y;
        proc rank; var y;
        proc rank groups = 2; var y2;
        proc rank groups = 2; var y3;
data block_a2;   set block_a;
        y4 = avg (y2, mod [y3+1,2]);   *
```

Under these instructions SAS will construct another data set called "block_a" which contains all of those cases from file one, for which variable block contains 'a'; i.e., block_a will contain the observations from block A of the experiment. The first call to the procedure ranks will cause one copy of the

---

* avg and mod are SAS built-in functions.

original data values to be replaced by their ranks within block A.

The second call to the rank procedure will cause another copy of the original data values, in variable y2, to be replaced with 0's and 1's; 0's for the values below the median and 1's for those at or above. The final call to the rank procedure does the same to the additive inverses of the original data values. Finally, in data set block_2a the original data values will be represented in variable y4 by 0's, 1's and ½'s; 0's for values below the median, 1's for values above it, and ½'s for values equal to the within-block median. The variables y and y4 will subsequently be used by the means procedure in calculations of the sums of within-block ranks by treatments and of the sums of indicator variable values by treatments, the values required in the Friedman and median test statistics respectively.

SAS outputs the following sums of values in variables y and y4, by treatments:

Treatment	1	2	3	4	5	6
y	22	22	15	6	6	13
y4	4	4	3	0	0	1

In this example, (1) is

$$S = \frac{12}{4 \times 6 \times 7} (22^2 + 22^2 + 15^2 + 6^2 + 6^2 + 13^2) \cong 18.4 \quad .$$

and (2) is

$$Q = 4^2 + 4^2 + 3^2 + 0^2 + 0^2 + 1^2 - \tfrac{1}{4} \times 4^2 \times 6 = 18 \quad .$$

Evoking more informative AØV tables from SAS: Analyses of data under the linear model and for which the design matrix is not of full rank are almost always done by methods which are equivalent to reparameterizing (as necessary and in one way or another) such that the design matrix does have full rank.

One might choose the following initial parameterization for the data in Bartlett (47):

yield equation: $\sqrt{Y_{ij}} = \mu + b_i + t_j + bt_{ij} + \epsilon_{ij}$

error structure: $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$

b_i's being the 4 block effects,

t_j's being the 6 treatment effects,

and bt_{ij}'s being the 24 block-treatment interactions, which happen to be completely confounded with error.

Of course, this model is not of full rank.

Let us pretend that the b treatments are b equally spaced levels of a single herbicide (rather than b types of herbicide as in the actual experiment) so that it may make more sense to speak of "linear effect of herbicide", "quadratic effect after linear", etc. Then the 5 d.f. for treatments may be spanned by the usual 5 orthogonal polynomials on 6 treatments:

linear ignoring quadratic, etc.:	5	3	1	-1	-3	-5 ;
quadratic after linear, ignoring cubic, etc. :	5	-1	-4	-4	-1	5 ;
cubic after linear and quadratic, ignoring quartic, etc. :	5	-7	-4	4	7	-5 ;
quartic after linear, quadratic and cubic, ignoring quintic :	1	-3	2	2	-3	1 ;
quintic after others :	1	-5	10	-10	5	-1 .

Many other sets of contrasts might be chosen; the ones above were chosen only in the interests of easy illustration. For instance, since treatment 1 was a control one might want to form the contrast of treatment 1 versus one or more of the other treatments to account for 1 of the 5 treatment d.f.'s; then one might choose 4 more which span the remaining d.f.'s.

Let us suppose further that the contrast-block interaction sums of squares are desired. One could choose block contrasts on grounds analogous to those on which treatment contrasts are chosen; however, perhaps most often they will be chosen just to span the aggregate d.f. for blocks with no attention being paid to orthogonality of the contrasts. If the treatment and block contrasts span the treatment and block d.f. respectively then the products of all treatment contrasts with all block contrasts will span the interaction d.f. Moreover, the total sum of squares associated with the products of a given treatment contrast with each of the block contrasts is the sum of squares, for the interaction of that contrast with blocks.

One may also remark that the sums of squares for the treatment contrasts mentioned earlier may also be calculated using SAS in another way. One has only to regress the dependent variable on treatment level, square of treatment level, cube of treatment level, etc. However, one cannot recover the sums of squares due to the interactions of contrasts with blocks.

Actual implementation of these principles in SAS is illustrated in the following program:

```
DATA ONE; INPUT BLOCK $ 1 Y_1 2-4
                          Y_2 5-7
                          Y_3 8-10
                          Y_4 11-13
                          Y_5 14-16
                          Y_6 17-19;
TREAT=1; Y=Y_1; OUTPUT;
TREAT=2; Y=Y_2; OUTPUT;
TREAT=3; Y=Y_3; OUTPUT;
TREAT=4; Y=Y_4; OUTPUT;
TREAT=5; Y=Y_5; OUTPUT;
TREAT=6; Y=Y_6; OUTPUT;
CARDS;
A438538 77 17 18115
B442422 61 51 26 57
C319377157 87 77100
D380315 52 16 20 45
DATA TWO; SET ONE;
COMMENT APPLY SQUARE ROOT TRANSFORMATION TO DATA ;
        Y=SQRT(Y);
COMMENT ALLOW FOR EXPLICIT CALCULATION OF MEAN SS;
        MEAN=1;
COMMENT CALCULATE TREATMENT CONTRAST DUMMY VARIABLES FOR "LINEAR",
        "QUADRATIC AFTER LINEAR", ETC;
LINEAR=5*((TREAT=1)-(TREAT=6))
      +3*((TREAT=2)-(TREAT=5))
        +(TREAT=3)-(TREAT=4);
QUADRATC=5*((TREAT=1)+(TREAT=6))
          -((TREAT=2)+(TREAT=5))
         -4*((TREAT=3)+(TREAT=4));
CUBIC=5*((TREAT=1)-(TREAT=6))
     -7*((TREAT=2)-(TREAT=5))
     -4*((TREAT=3)-(TREAT=4));
QUARTIC=(TREAT=1)+(TREAT=6)
     -3*((TREAT=2)+(TREAT=5))
     +2*((TREAT=3)+(TREAT=4));
QUINTIC=(TREAT=1)-(TREAT=6)
     -5*((TREAT=2)-(TREAT=5))
    +10*((TREAT=3)-(TREAT=4));
COMMENT  CALCULATE SQUARES OF TREATMENTS LEVELS, CUBES, ETC;
         SQUARE =TREAT*TREAT;
         CUBE=SQUARE*TREAT;
         HYPER1=CUBE*TREAT;
         HYPER2=HYPER1*TREAT;
COMMENT CALCULATE BLOCK CONTRAST DUMMY VARIABLE VALUES;
```

```
46                  B_CON_1=(BLOCK='A')-(BLOCK='B');
47                  B_CON_2=(BLOCK='B')-(BLOCK='C');
48                  B_CON_3=(BLOCK='C')-(BLOCK='D');
49          COMMENT CALCULATE BLOCK CONTRAST X TREAT CONTRAST DUMMY VARIABLE VALUES;
50                  I_1_L=LINEAR*B_CON_1;
51                  I_2_L=LINEAR*B_CON_2;
52                  I_3_L=LINEAR*B_CON_3;
53                  I_1_QD=QUADRATC*B_CON_1;
54                  I_2_QD=QUADRATC*B_CON_2;
55                  I_3_QD=QUADRATC*B_CON_3;
56                  I_1_C=CUBIC*B_CON_1;
57                  I_2_C=CUBIC*B_CON_2;
58                  I_3_C=CUBIC*B_CON_3;
59                  I_1_QR=QUARTIC*B_CON_1;
60                  I_2_QR=QUARTIC*B_CON_2;
61                  I_3_QR=QUARTIC*B_CON_3;
62                  I_1_QN=QUINTIC*B_CON_1;
63                  I_2_QN=QUINTIC*B_CON_2;
64                  I_3_QN=QUINTIC*B_CON_3;
65          PROC REGR; CLASSES TREAT BLOCK;
66          COMMENT SS'S FOR MEAN, TREATMENT, BLOCKS, RESIDUAL;
67                   MODEL Y=MEAN BLOCK TREAT / NOINT;
68          COMMENT SS'S DUE TO INDIVIDUAL TREATMENT CONTRASTS AND TO INTERACTIONS BETWEEN
69                  THESE CONTRASTS AND BLOCKS;
70                   MODEL Y=BLOCK LINEAR QUADRATC CUBIC QUARTIC QUINTIC
71                                 I_1_L I_2_L I_3_L  I_1_QD I_2_QD I_3_QD
72                                 I_1_C I_2_C I_3_C  I_1_QR I_2_QR I_3_QR
73                                 I_1_QN I_2_QN I_3_QN  ;
74          PROC REGR; CLASSES BLOCK;
75          COMMENT SS'S DUE TO "LINEAR", "QUADRATC AFTER LINEAR", ETC BY OTHER METHOD;
76                   MODEL Y=BLOCK TREAT SQUARE CUBE HYPER1 HYPER2;
```

Program explication: The first 13 lines of the program cause the data to be read and stored as in the 1st example. The construction of the data set which is actually operated upon to form the AØV commences with line 18. Let us look at this construction in slightly more detail:

1. One can cause SAS to print the correction for the mean sums of squares by making the mean an explicit part of the model as in line 67, forcing SAS not to correct for the mean on its own by using the "NØINT" parameter. Then line 22 is meant merely to construct a variable for the intercept in the model of line 67.

2. Lines 25 through 39 cause SAS to construct the variables, the "contrasts", representing "linear effect of treatments", "quadratic after linear", etc. upon which the dependent variable is regressed in line 70 to obtain sums of squares for these contrasts.

3. Lines 41 to 44 cause the construction of the variables used for the "other" method of calculation of sums of squares due to "linear", "quadratic after linear", etc., in line 76.

4. Lines 46, 47 and 48 cause construction of 3 contrasts which span the block d.f.

5. The 3 d.f. spanned by the variables constructed in lines 50, 51 and 52 are associated with the interaction of the "linear" contrast with blocks. Likewise, the next three are for interaction of "quadratic after linear"; and so on, to line 64.

Actual calculations of the AØV tables are made with two invocations of the regr procedure and 3 model statements in the final section of the program.

Each model statement results in the output of a single AØV table. One can consolidate information from this output as in the following AØV table:

Source	df	SS
CFM	1	3199.898
Blocks	3	29.617
Treatments	5	904.624
Linear (L)	1	599.979
Quadratic after L (qd)	1	158.225
Cubic after L, qd (c)	1	79.869
Quartic after L, qd, c (qr)	1	53.621
Quintic after L, qd, c, qr (qn)	1	12.930
Blocks x treatments	15	52.860
L x blocks	3	15.560
qd x blocks	3	20.190
c x blocks	3	12.219
qr x blocks	3	2.768
qn x blocks	3	2.122
Total	24	4187

The total SS and correction for the mean have been taken from the 1st AØV table output by SAS. The blocks SS might have come from any of the three tables' output. The treatments SS was extracted from the first table. The SS's due to "linear (L)", "quadratic after L (qd)", etc. might have been obtained from either of the last two tables.

SAS outputs the SS corresponding to each d.f. for the variables spanning the 15 interaction d.f. The SS for the 3 d.f. for blocks x linear was obtained as the sum of the SS's for linear x 1st block contrast, linear x 2nd block contrast and linear x 3rd block contrast. The SS for blocks x qd, etc. were obtained in a similar fashion. The total blocks x treatments SS was available labeled "error" in the table SAS output 1st.

Tukey's (49) single degree of freedom: It is straightforward to implement the method for calculation of Tukey's (49) single degree of freedom for nonadditivity outlined in Snedecor and Cochran (67). We illustrate by showing how the calculations for this sum of squares may be facilitated for the data in table 2:

```
001                 DATA ONE; INPUT WEEK 1
002                                 SYMBOL_1 $ 2  YIELD_1 3-7 2
003                                 SYMBOL_2 $ 8  YIELD_2 9-13 2
004                                 SYMBOL_3 $ 14 YIELD_3 15-19 2
005                                 SYMBOL_4 $ 20 YIELD_4 21-25 2;
006                 SYMBOL=SYMBOL_1; YIELD=YIELD_1; COW=1; OUTPUT;
007                 SYMBOL=SYMBOL_2; YIELD=YIELD_2; COW=2; OUTPUT;
008                 SYMBOL=SYMBOL_3; YIELD=YIELD_3; COW=3; OUTPUT;
009                 SYMBOL=SYMBOL_4; YIELD=YIELD_4; COW=4; OUTPUT;    CARDS;
010                 1D 8375C 3325A 4575B14450
011                 2A 4700D 4400B 3675C13575
012                 3C 3650B 4150D 2750A12550
013                 4B 2400A 3875C 3050D12425
014                 PROC REGR; CLASSES SYMBOL COW WEEK;
015                     MODEL YIELD=SYMBOL COW WEEK;
016                           OUTPUT OUT=TWO RESIDUAL Y_RES PREDICTED Y_HAT;
017                 DATA THREE; SET TWO; Y_HAT_2=Y_HAT*Y_HAT; PROD=Y_RES*Y_HAT_2;
018                 PROC MEANS; VAR PROD;
019                 PROC REGR; CLASSES SYMBOL COW WEEK;
020                     MODEL Y_HAT_2=SYMBOL COW WEEK;
```

Program explication: The 1st call to the regr procedure in line 14 causes the calculation of the usual AØV and also of the residuals and predicted values. The squares of the predicted values and the products of these squares and the residuals are then calculated in line 17. The call to means in line 18 results in the calculation of, among other things, the sum of these products.* Calculation of the error SS for the squares** is accomplished upon the call to regr in line 19.

* N in Snedecor and Cochran's notation.

** D in Snedecor and Cochran's notation.

The SS for Tukey's (49) single d.f. can be calculated as in Snedecor and Cochran (67) as

$$\frac{18803.49^2}{3085284} \cong 114.6$$

where

18803.49 is the sum of the products of the residuals and squares of predicted values output by the procedure means,

and

3085284 is the error SS for the squares of the predicted values, output by the regr procedure at its final invocation.

Cox's (58) analysis: Referring to Cox's data in table 2: it may be expected that the yield curves for the four cows over weeks may not be parallel so that Cox's analysis involving the following yield equation may be appropriate:

$$y_{ij} = \mu + c_j + k_j r_{ij} + t_i + \epsilon_{ij}$$

where

$\mu \equiv$ the general mean parameter

c_j and $k_j \equiv$ intercept and slope parameters for linear change in jth cow over weeks

$t_i \equiv$ ith treatment parameter

$r_{ij} \equiv$ row constant for ith treatment in the jth column.

The following program will cause SAS to calculate SS's due to treatments, rows, cows and individual "within-cow" regressions:

```
01                DATA ONE; INPUT WEEK 1
02                              SYMBOL_1 $ 2 YIELD_1 3-7   2
03                              SYMBOL_2 $ 8 YIELD_2 9-13  2
04                              SYMBOL_3 $ 14 YIELD_3 15-19 2
05                              SYMBOL_4 $ 20 YIELD_4 21-25 2;
06                SYMBOL=SYMBOL_1; YIELD=YIELD_1; YIELD2=YIELD*YIELD; COW=1; OUTPUT;
07                SYMBOL=SYMBOL_2; YIELD=YIELD_2; YIELD2=YIELD*YIELD; COW=2; OUTPUT;
08                SYMBOL=SYMBOL_3; YIELD=YIELD_3; YIELD2=YIELD*YIELD; COW=3; OUTPUT;
09                SYMBOL=SYMBOL_4; YIELD=YIELD_4; YIELD2=YIELD*YIELD; COW=4; OUTPUT;
10                CARDS;
11                1D 8375C 5325A 4575B14450
12                2A 4700D 4400B 3675C13575
13                3C 3650B 4150D 2750A12550
14                4B 2400A 3875C 3050D12425
15                DATA FIVE;   SET ONE;
16                COMMENT FORM DUMMY REGRESSOR VARIABLE VALUES FOR COX'S ANALYSIS FOR 1ST COW;
17                REGN_1=0;
18                IF WEEK=1 AND COW=1 THEN REGN_1 =-3;
19                IF WEEK=2 AND COW=1 THEN REGN_1 =-1;
20                IF WEEK=3 AND COW=1 THEN REGN_1 = 1;
21                IF WEEK=4 AND COW=1 THEN REGN_1 = 3;
22                COMMENT SIMILARLY FOR OTHERS;
23                REGN_2=0; REGN_3=0; REGN_4=0;
24                IF WEEK=1 AND COW=2 THEN REGN_2=-3;
25                IF WEEK=2 AND COW=2 THEN REGN_2=-1;
26                IF WEEK=3 AND COW=2 THEN REGN_2= 1;
27                IF WEEK=4 AND COW=2 THEN REGN_2= 3;
28                IF WEEK=1 AND COW=3 THEN REGN_3=-3;
29                IF WEEK=2 AND COW=3 THEN REGN_3=-1;
30                IF WEEK=3 AND COW=3 THEN REGN_3= 1;
31                IF WEEK=4 AND COW=3 THEN REGN_3= 3;
32                IF WEEK=1 AND COW=4 THEN REGN_4=-3;
33                IF WEEK=2 AND COW=4 THEN REGN_4=-1;
34                IF WEEK=3 AND COW=4 THEN REGN_4= 1;
35                IF WEEK=4 AND COW=4 THEN REGN_4= 3;
36                PROC REGR; CLASSES COW  SYMBOL;
37                    MODEL YIELD=COW REGN_1 REGN_2 REGN_3 REGN_4 SYMBOL;
```

Programm explication: The reader may recognize the integers at the ends of lines 18 through 21 as the most common representation of the "linear ignoring higher degree effects" orthogonal polynomial over 5 equally spaced treatment levels. Thus, the variable regn_1 corresponds to a d.f. for a regression of yields for cow number 1 over the 4 weeks. Variables regn_2, regn_3 and regn_4 are defined analogously.

SAS outputs SS's for each d.f. corresponding to a "within-cow" slope; i.e. for each of the variables regn_1 through regn_4. These 4 SS's may be summed to form the SS due to regression. Then this and the other information from the output can be consolidated as in the following table:

Source	d.f.	SS
Cows	3	25576.355
Regressions (ignoring treatments)	4	2213.353
Treatments (after regressions)	3	106.643
Residual	5	197.676
Corrected Total	15	28094.027

Examination of residuals: Anscombe and Tukey (63) suggest a number of ways in which residuals may be examined with a mind to checking assumptions made, including those concerning error structure. Roughly speaking, one procedure suggested is to plot the elements of the order statistic associated with the residuals against the ordered population medians of the elements of the order statistics of the same length for a simple random sample from the N(0,1) population. It is fairly easy to program SAS to do this, thus obviating considerable tedium.

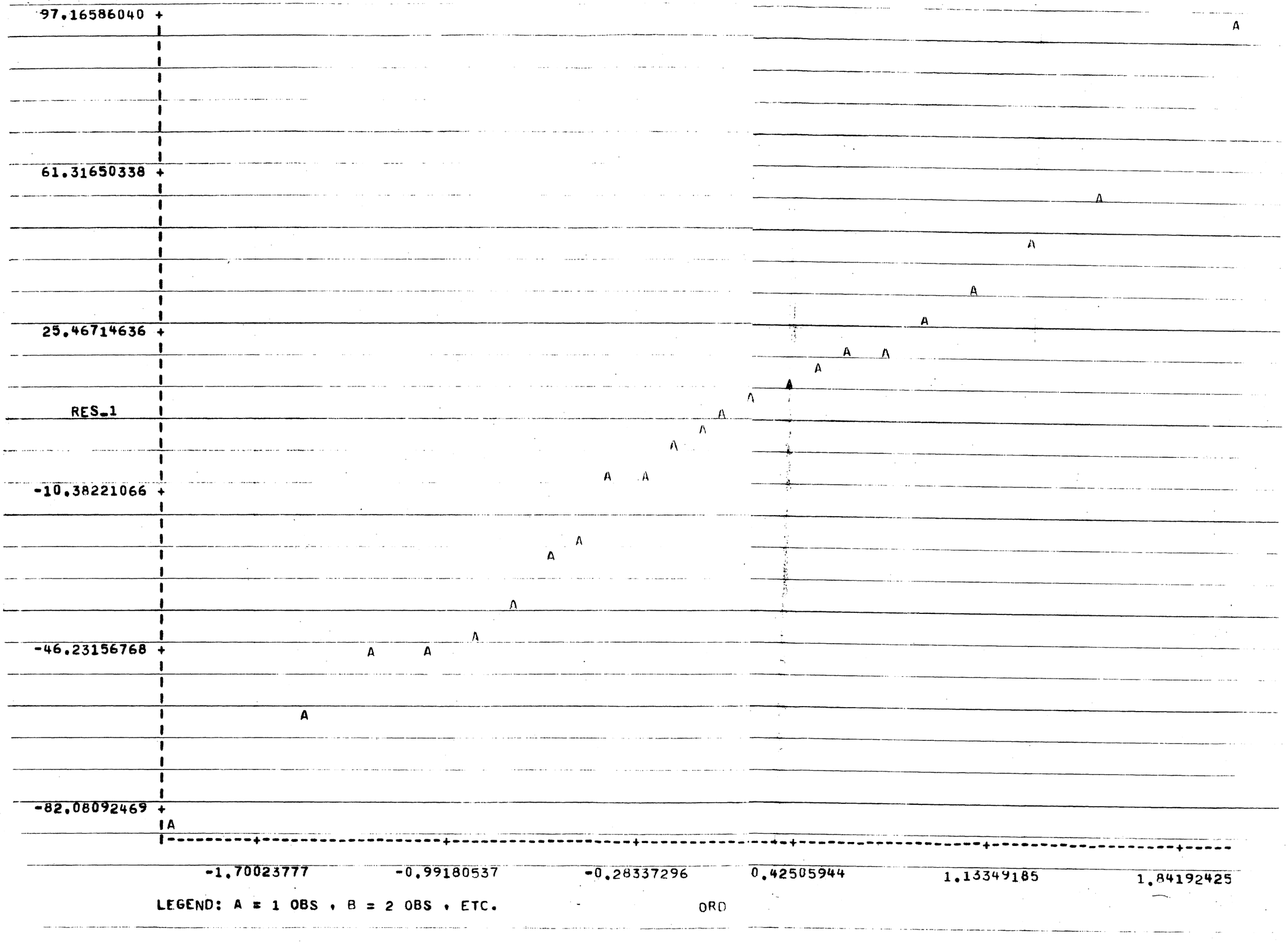
We use the data from table 1 to illustrate:

```
01          DATA ONE; INPUT BLOCK $ 1 Y_1 2-4
02                                    Y_2 5-7
03                                    Y_3 8-10
04                                    Y_4 11-13
05                                    Y_5 14-16
06                                    Y_6 17-19;
07          TREAT=1; Y=Y_1; OUTPUT;
08          TREAT=2; Y=Y_2; OUTPUT;
09          TREAT=3; Y=Y_3; OUTPUT;
10          TREAT=4; Y=Y_4; OUTPUT;
11          TREAT=5; Y=Y_5; OUTPUT;
12          TREAT=6; Y=Y_6; OUTPUT;
13          DROP Y_1 Y_2 Y_3 Y_4 Y_5 Y_6;
14
15          CARDS;
16          A438538 77 17 18115
17          B442422 61 31 26 57
18          C519377157 87 77100
19          D380315 52 16 20 45
20          PROC REGR; CLASSES BLOCK TREAT;
21              MODEL Y=BLOCK TREAT;
22                    OUTPUT OUT=TWO RESIDUAL RES_1;
23          PROC SORT; BY RES_1;
24          DATA FIRST_24; INPUT;
25                    I=1;
26            LOOP: ORD=PROBIT((I-.5)/24); OUTPUT;
27                  I=I+1; IF I<=24 THEN GOTO LOOP;
28          DROP I; CARDS;
29
30          DATA THREE; MERGE TWO   FIRST_24;
31          DROP BLOCK TREAT Y;
32          PROC PLOT;
```

Program explication: Lines 21 and 22 cause the residuals, variable res_1, to be written to file two, having corrected for mean, blocks and treatments. Lines 24 to 29 direct the calculation of the required population medians. Line 30 writes the median-residual pairs to file three. The pairs are plotted in line 32.

This program results in output which includes the following graph:

PLOT OF RES_1 VS ORD



LEGEND: A = 1 OBS , B = 2 OBS , ETC.

Construction of certain series of orthogonal polynomials: The reader might refer to Robson and Atkinson (58). The technique they propose involves the weighted regression of within-treatment slopes on orthogonal polynomial functions of the adjusted treatment means; as you may presently observe, the weights are unequal and spacings of the independent variable are unequal. One approach to the calculation of the SS's due to each of these polynomials using SAS would be to begin by generating the polynomial values evaluated at the values assumed by the independent variable. Draper and Smith (66) mention a method for the case of equal weighting which is easily adapted to unequal weights using the simple technique mentioned in Bell (74).

The weights, slopes and means for the example mentioned in Robson and Atkinson are given in the following table:

i	weights w_i	within treatment slopes b_i	adjusted treatment means α_i
1	1610	.00978881	1.26575
2	640.5	.0136534	1.11043
3	1066.1	.00525936	.926964
4	945.6	.00123096	.807619

Generation of polynomial values and weighted regression of slopes on means may be accomplished by the following program:

```
001                 DATA ONE; INPUT WEIGHT 1-5 1
002                                 SLOPE 6-13 8
003                                 ADJMEAN 14-20 6;
004                     ROOT_W=SQRT(WEIGHT);
005                     WADJM_1=ROOT_W*ADJMEAN;
006                     WADJM_2=WADJM_1*ADJMEAN;
007                     WADJM_3=WADJM_2*ADJMEAN;
008                     WSLOPE=ROOT_W*SLOPE;
009                 CARDS;
010                 16100009788811265750
011                  64050136534011104 30
012                 10661005259360926964
013                  94560012309608076 19
015                 PROC REGR;
016                     MODEL WADJM_1=ROOT_W / NOINT;
017                          OUTPUT OUT=TWO RESIDUAL LINEAR;
019                 PROC REGR;
020                     MODEL WADJM_2=ROOT_W LINEAR / NOINT;
021                          OUTPUT OUT=THREE RESIDUAL QUAD;
023                 PROC REGR;
024                     MODEL WADJM_3=ROOT_W LINEAR QUAD / NOINT;
025                          OUTPUT OUT=FOUR RESIDUAL CUB;
026                 PROC PRINT;
027                 PROC REGR;
028                     MODEL WSLOPE =ROOT_W LINEAR QUAD CUB / NOINT;
032                 PROC REGR;
033                     MODEL WSLOPE=ROOT_W WADJM_1 WADJM_2 WADJM_3 / NOINT;
```

Program explication: Lines 1 through 13 of the program store the weights, slopes and adjusted means from the data cards on file "one", along with versions of these values which are used in calculating a number of weighted regressions by the principle mentioned in Bell (74).

The 1st call to the regr procedure results in the calculation of the weighted regression of the adjusted means on the constant 1; the residuals, in variable linear, are written to file two which is processed by regr at its next invocation. This call results in the weighted regression of the squares of the adjusted means on the constant 1 and the variable linear. Thus, the 1st three invocations of regr result in the calculation of values of the three orthogonal polynomials corresponding to "linear effect" of adjusted means, "quadratic after linear", and "cubic after linear and quadratic", for a weighted regression. These values are made available in file four and the 4th call to regr results in the weighted regression of the slopes on the constant 1 and the three orthogonal polynomials. (The final call to regr points up the fact that Robson and Atkinson's AØV can be calculated rather more directly.)

The orthogonal polynomial values may be read from the print-out of data set four.

The SS's corresponding to the common slope coefficient and to the orthogonal polynomials may be read from the print-out corresponding to the 4th regression. The common regression coefficient, the slope estimate associated with the covariate, may also be read here as the "b value" corresponding to root_w (this last value happens to be .007338). Then the results may be consolidated as in the following table——for the regression of slopes on adjusted means:

Source	d.f.	SS
Total	4	.30459
Mean (covariate)	1	.22950
Linear	1	.05149
Quadratic after linear	1	.01677
Cubic after linear and quadratic	1	.00683

Acknowledgements

My thanks are due to Dr. F. Cady for suggestions toward improved exposition and for a careful, critical discussion of the mimeo. May I also thank Messrs. D. Hall, Y. Ahn and D. Schuirmann for reading the mimeograph, and Ms. Norma Phalen for the typing.

Thanks (belatedly) to Dr. D. Solomon and Mr. R. Nair for examining the preceding mimeo in this series.

(Dr. F. Cady has pointed out to me the fact that S,(1), could be much more easily calculated (Hollander and Wolfe (73), p. 143) from the SS for treatments on the ranks, a possibility I happily ignored in my state of childish euphoria, manipulating SAS.)

References

Anscombe and Tukey (63) "Examination of residuals", Technometrics 5, p. 141.

Bell (74) "Spending a rainy afternoon with SAS—fitting weighted linear regressions with SAS", BU-527-M in the Biometrics Unit Mimeo Series, Cornell University.

Cox (58) "The analysis of Latin square designs with individual curvatures in one direction", JRSS B 20, p. 193-204.

Draper and Smith (66) Applied Regression Analysis, John Wiley, N.Y., p. 156.

Federer (74) Communications to students of Course Statistics 513 at Cornell University, Ithaca, N. Y.

Hájek and Sidák (67) Theory of Rank Tests, Academic Press, N. Y., p. 117.

Hollander and Wolfe (73) Nonparametric Statistical Methods, Wiley, N.Y., p. 139.

Robson and Atkinson (58) "Testing homogeneity of regression coefficients in a one-way analysis of variance", Biometrics 16, p. 593.

Snedecor and Cochran (67) Statistical Methods, Iowa State University, Ames, Iowa, p. 331.

Tukey (49) "One degree of freedom for non-additivity", Biometrics 5, p. 232.